DD2434 – Advanced Machine Learning

Lecture 6: Representation Learning

Pawel Herman

Computational Science and Technology (CST)

KTH Royal Institute of Technology

Short outline for today

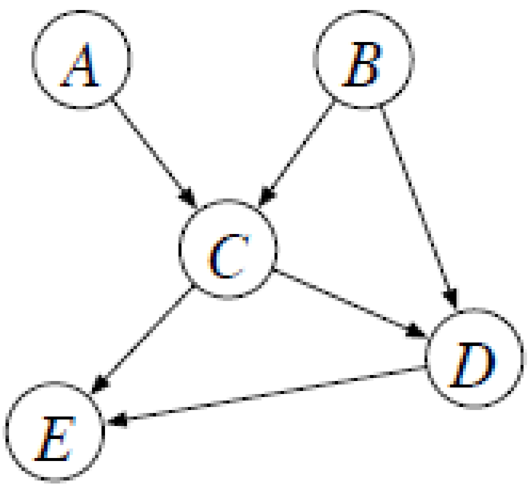
1. Basics of graphical models

2. Factor analysis with Probabilistic Principal Component Analysis (PPCA)

Graphical models – basics

"Marriage between probability theory and graph theory"

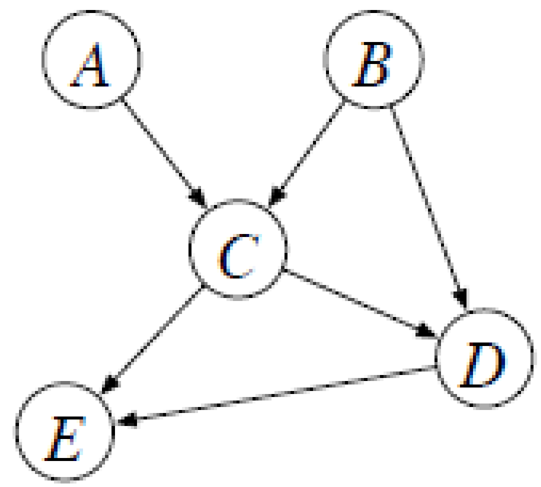
- tool for dealing with uncertainty, independence, and complexity
- an intuitive way of representing and visualising the relationships between many variables.



Graphical models – basics

"Marriage between probability theory and graph theory"

- tool for dealing with uncertainty, independence, and complexity
- an intuitive way of representing and visualising the relationships between many variables.



Focus on directed acyclic graphs (DAGs)

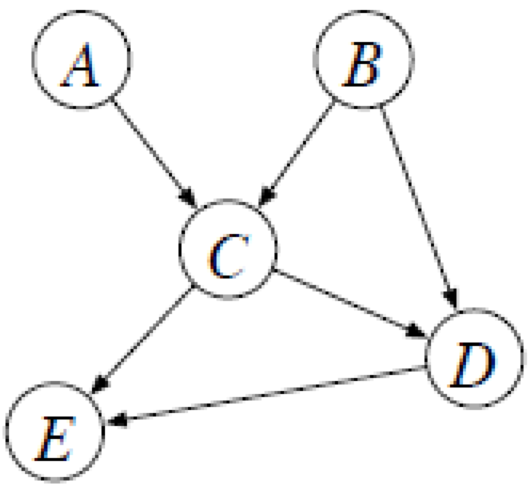
- factorisation of the joint probability distribution

$$p(A,B,C,D,E) = p(A)p(B)p(C|A,B)p(D|B,C)p(E|C,D)$$

Graphical models – basics

"Marriage between probability theory and graph theory"

- tool for dealing with uncertainty, independence, and complexity
- an intuitive way of representing and visualising the relationships between many variables.



Focus on directed acyclic graphs (DAGs)

- factorisation of the joint probability distribution *(associated with the whole graph)*

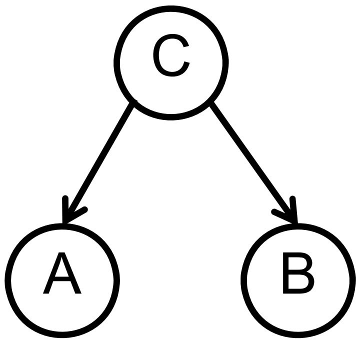
$$p(X_1, \ldots, X_n) = \prod_{i=1}^{n} p(X_i | X_{\mathrm{pa}(i)})$$

Graphical models – basics

Conditional independence relationships

- way to abstract out such relationships between variables from the details of their parametric forms

"Is A dependent on B given that we know the value of C?"



Graphical models – basics

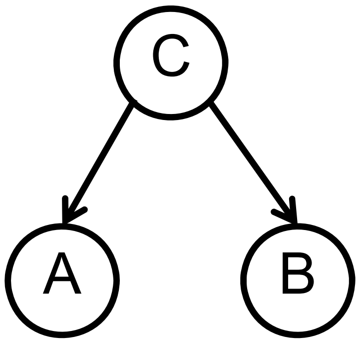
Conditional independence relationships

➢ way to abstract out such relationships between variables from the details of their parametric forms

"Is A dependent on B given that we know the value of C?"

A and B are conditionally probabilistically independent given C if and only if

$$p(A,B|C) = p(A|C) * p(B|C)$$



No direct link between children nodes, A and B

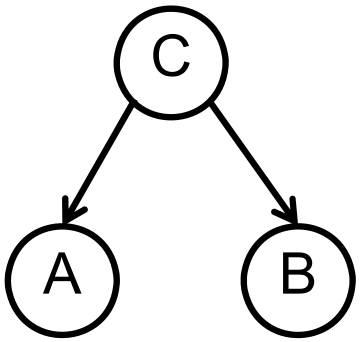
Graphical models – basics

Conditional independence relationships

- way to abstract out such relationships between variables from the details of their parametric forms

"Is A dependent on B given that we know the value of C?"

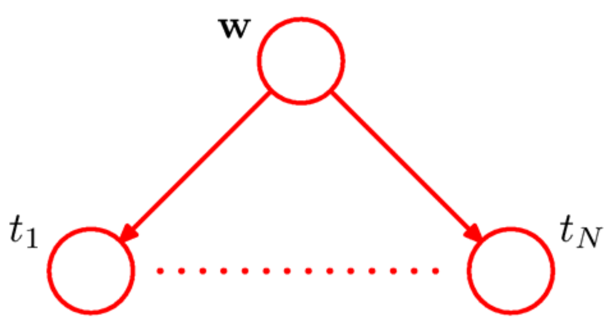
A and B are conditionally probabilistically independent given C if and only if



$$p(A,B|C) = p(A|C) * p(B|C)$$

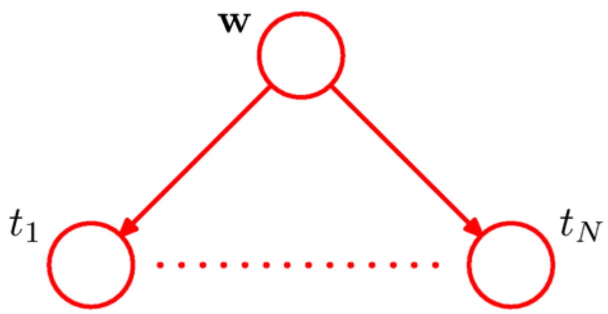
$$p(A,B,C) = p(A|C) * p(B|C)*p(C)$$

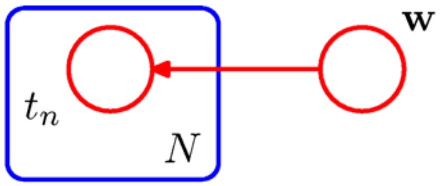
Graphical models – linear regression example



$$p(\mathbf{t}, \mathbf{w}) = p(\mathbf{w}) \prod_{n=1}^{N} p(t_n \mid \mathbf{w})$$

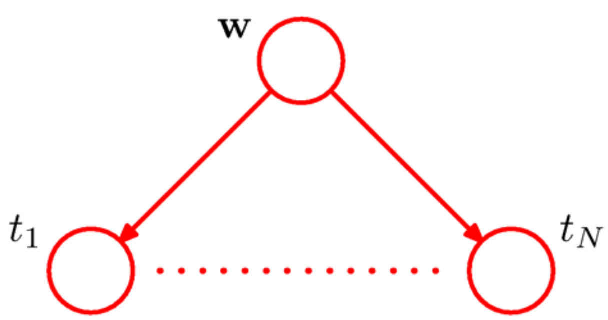
Graphical models – linear regression example

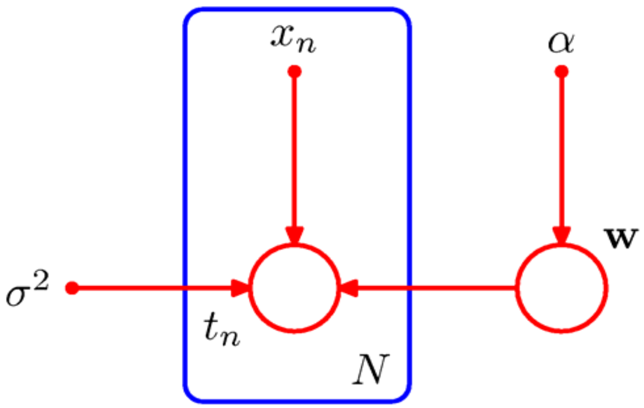




$$p(\mathbf{t},\mathbf{w}) = p(\mathbf{w})\prod_{n=1}^{N} p(t_n \mid \mathbf{w})$$

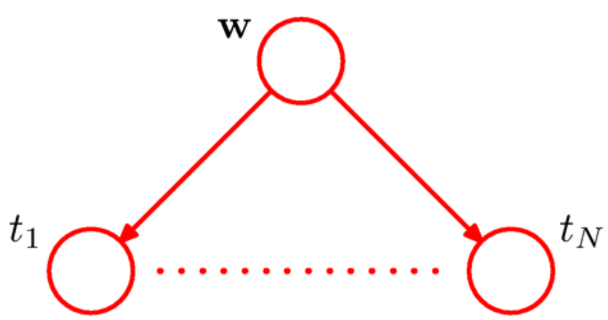
Graphical models – linear regression example

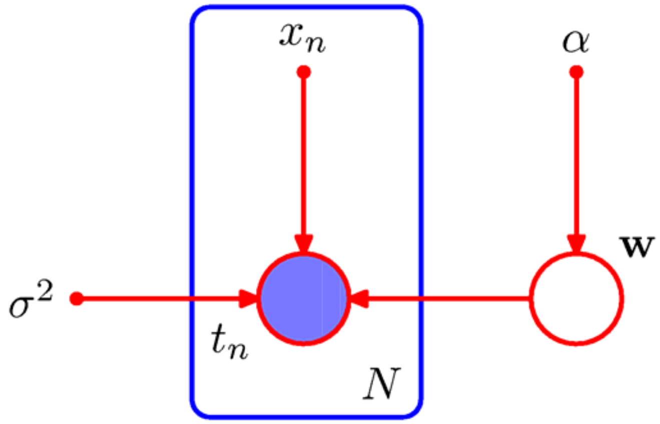




$$p(\mathbf{t},\mathbf{w}\,|\,\mathbf{X},\alpha,\sigma^2) = p(\mathbf{w}\,|\,\alpha)\prod_{n=1}^{N} p(t_n\,|\,\mathbf{w},x_n,\sigma^2)$$

Graphical models – linear regression example



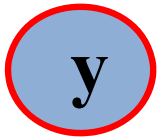


$$p(\mathbf{t},\mathbf{w}\,|\,\mathbf{X},\alpha,\sigma^2) = p(\mathbf{w}\,|\,\alpha)\prod_{n=1}^{N} p(t_n\,|\,\mathbf{w},x_n,\sigma^2)$$

Solid Blue Circle means observable values.

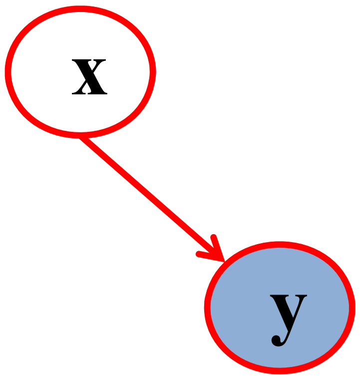
Latent variable models

- observable data
- task to model $p(y)$



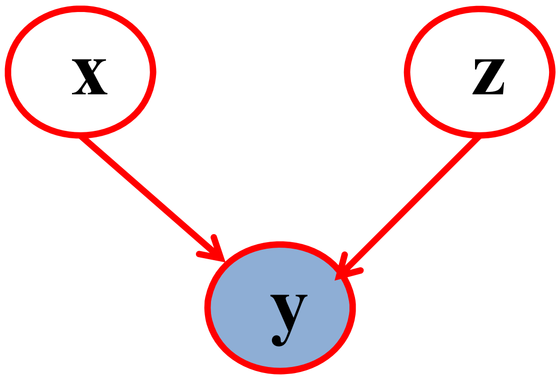
Latent variable models

- observable data
- task to model $p(y)$
- unobservable: *latent* variables



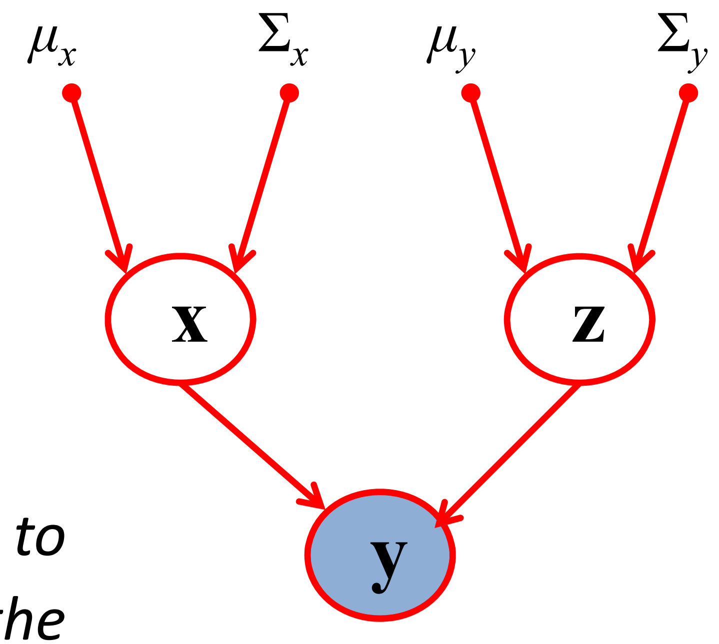
Latent variable models

- observable data
- task to model $p(y)$
- unobservable: *latent* variables



Latent variable models

- observable data
- task to model $p(y)$
- unobservable: *latent* variables

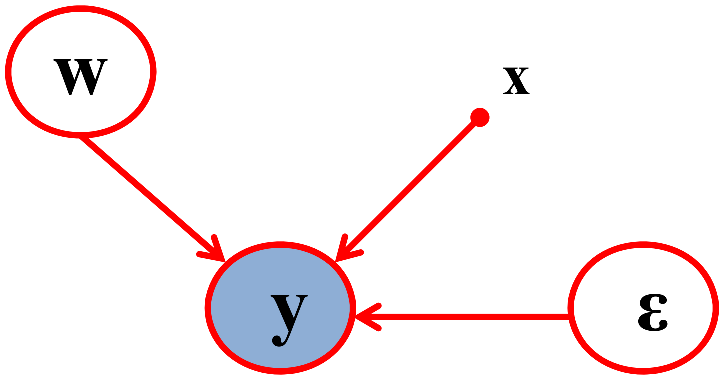


*"The primary role of the **latent variables** is to allow a complicated distribution over the observed variables to be represented in terms of a model constructed from simpler (typically exponential family) conditional distributions."*

Bishop 2006

Latent variable models

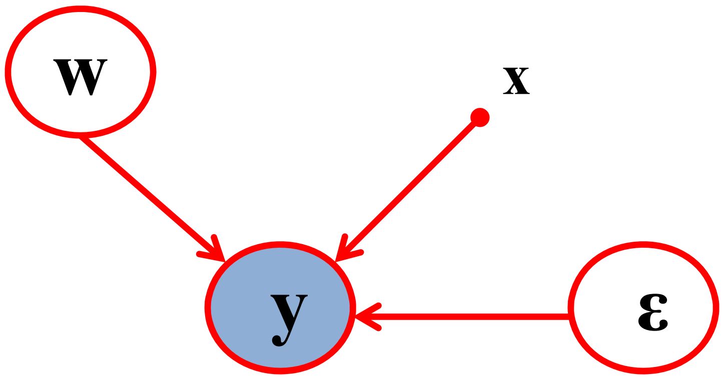
- observable data
- task to model $p(y)$
- unobservable: *latent* variables



$$y_i = \mathbf{w}^{\mathrm{T}}\mathbf{x}_i + \varepsilon$$

Latent variable models

- observable data
- task to model $p(y)$
- unobservable: *latent* variables
- "explaining away" phenomenon



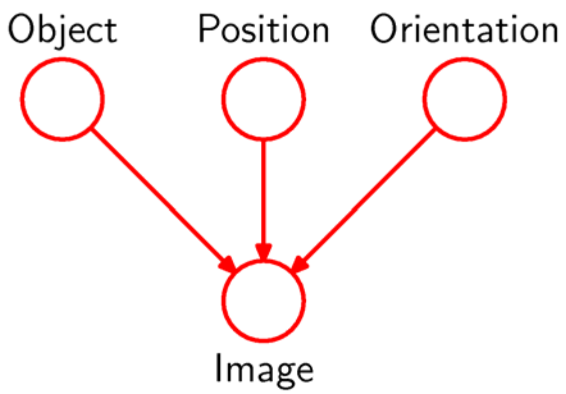
$$y_i = \mathbf{w}^{\mathrm{T}}\mathbf{x}_i + \varepsilon$$

Latent variable models

- observable data
- task to model $p(y)$
- unobservable: *latent* variables
- “explaining away” phenomenon
- **generative** models for conditional distributions being

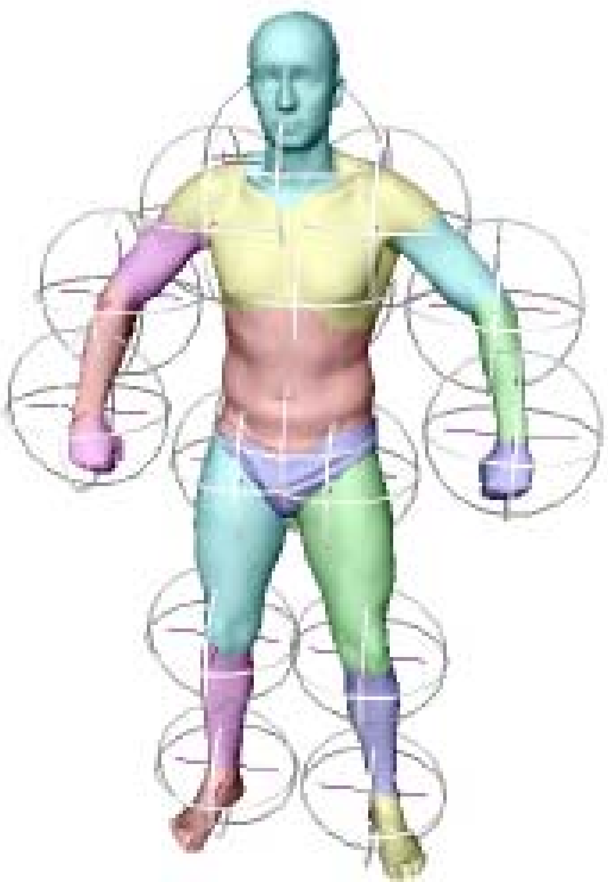
parameterised by a latent *random* variable - input: $p(y|x)$

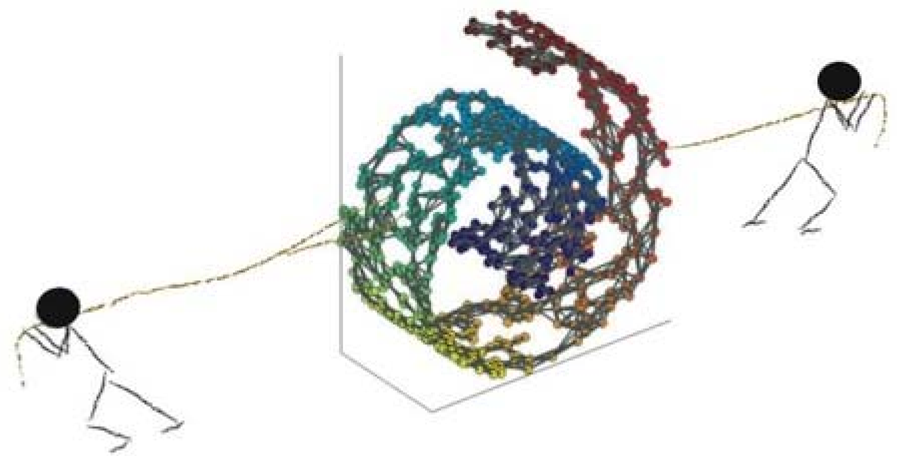
(the concept of *ancestral* sampling -> creation of the *“fantasy”* observed data)



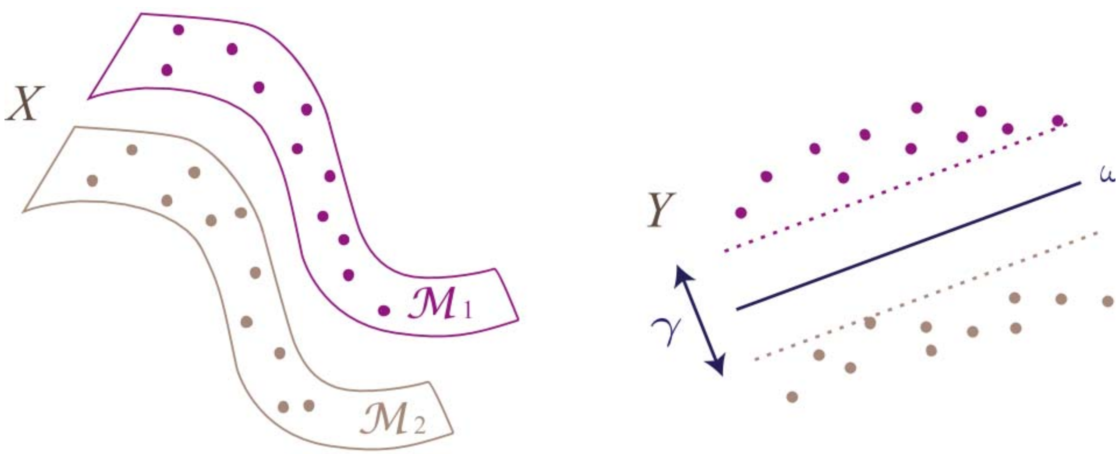
Bishop 2006

Data dimensionality





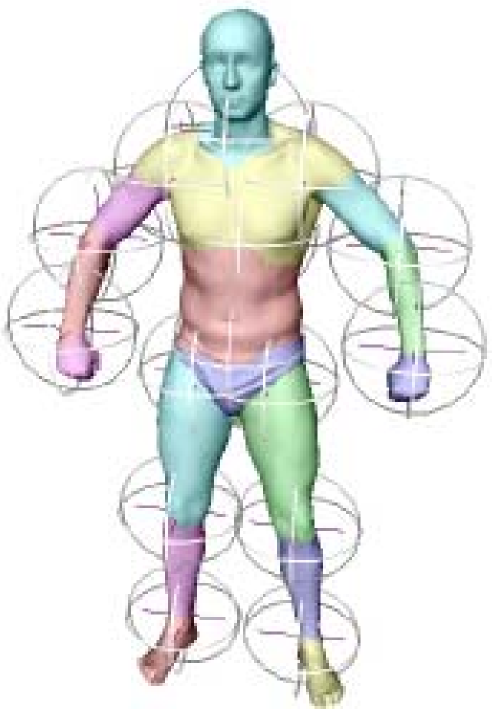
Kilian Q. Weinberger



adapted from Carl Henrik Ek's lecture (DD2434, 2015)

Intrinsic dimensionality

- Data dimensionality vs intrinsic dimensionality
  - fewer “data-driven” degrees of freedom
  - discovering lower-dimensional manifold
  - re-parameterisation of data
    - to understand/interpret relevant aspects of data
    - to boost generalisation
    - for computational efficiency



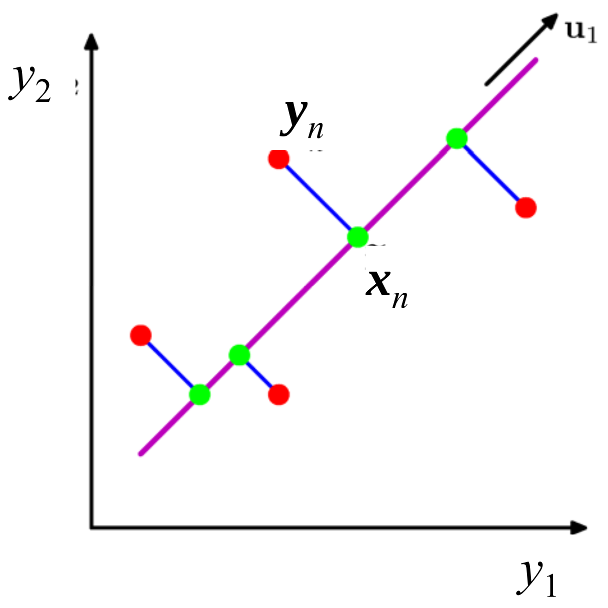
adapted from Carl Henrik Ek’s lecture (DD2434, 2015)

Popular PCA

- Principal component analysis (PCA)
  - Project D-dimensional data $\mathbf{Y}=\{\mathbf{y}_n\}$, on to the principal subspace $\mathbf{X}=\{\mathbf{x}_n\}$ of lower-dimensionality, $M<D$

$$\mathbf{X} = \mathbf{Y}\mathbf{u}$$

  - Variance of the projected data should be maximised



Popular PCA

- Principal component analysis (PCA)
  - Project D-dimensional data $\mathbf{Y}=\{\mathbf{y}_n\}$, on to the principal subspace $\mathbf{X}=\{\mathbf{x}_n\}$ of lower-dimensionality, $M<D$

$$\mathbf{X} = \mathbf{Y}\mathbf{u}$$

  - Variance of the projected data should be maximised

$$\max_{\mathbf{u}} \arg\{\mathrm{var}(\mathbf{X}) : \mathbf{X} = \mathbf{Y}\mathbf{u}\} = \max_{\mathbf{u}} \arg\left\{(\mathbf{Y}\mathbf{u})^{\mathrm{T}} \mathbf{Y}\mathbf{u}\right\}, \quad \text{subject to } \mathbf{u}^{\mathrm{T}}\mathbf{u} = 1$$

Popular PCA

- Principal component analysis (PCA)
  - Project D-dimensional data $\mathbf{Y}=\{\mathbf{y}_n\}$, on to the principal subspace $\mathbf{X}=\{\mathbf{x}_n\}$ of lower-dimensionality, $M<D$

$$\mathbf{X} = \mathbf{Y}\mathbf{u}$$

  - Variance of the projected data should be maximised

$$\max_{\mathbf{u}} \arg\{\mathrm{var}(\mathbf{X}) : \mathbf{X} = \mathbf{Y}\mathbf{u}\} = \max_{\mathbf{u}} \arg\left\{(\mathbf{Y}\mathbf{u})^{\mathrm{T}}\, \mathbf{Y}\mathbf{u}\right\}, \quad \text{subject to } \mathbf{u}^{\mathrm{T}}\mathbf{u} = 1$$

$$\Downarrow$$

$$\mathbf{u}_k^{\mathrm{T}} \,\mathrm{cov}(\mathbf{Y})\, \mathbf{u}_k = \lambda_k$$

eigenvector formulation

Latent variable model

- Data observed $\mathbf{Y}$

$$p(\mathbf{Y})$$

Latent variable model

- Data observed $\mathbf{Y}$

$$p(\mathbf{Y})$$

- Assumption: $\mathbf{Y} \in \mathbb{R}^{N \times D}$ have been generated from $\mathbf{X} \in \mathbb{R}^{N \times M}$ by means of some generative mapping, f

$$p(\mathbf{Y} \mid \mathrm{f}, \mathbf{X}), \text{ where } \mathrm{f} : \mathbf{X} \rightarrow \mathbf{Y}$$

Latent variable model

- Data observed $\mathbf{Y}$

$$p(\mathbf{Y})$$

- Assumption: $\mathbf{Y} \in \mathbb{R}^{N \times D}$ have been generated from $\mathbf{X} \in \mathbb{R}^{N \times M}$ by means of some generative mapping, f

$$p(\mathbf{Y} \mid \mathrm{f}, \mathbf{X}), \text{ where } \mathrm{f} : \mathbf{X} \rightarrow \mathbf{Y}$$

- $\mathbf{X}$ is then considered as *latent* variable

$$\mathbf{x}_n \xrightarrow{\mathrm{f}} \mathbf{y}_n$$

Linear latent variable model

$$p(\mathbf{Y} \mid \mathbf{W}, \mathbf{X}) = \prod_{n=1}^{N} p(\mathbf{y}_n \mid \mathbf{W}, \mathbf{x}_n)$$

$$p(\mathbf{y} \mid \mathbf{W}, \mathbf{x}) = \mathcal{N}\left(\mathbf{W}\mathbf{x} + \boldsymbol{\mu}, \boldsymbol{\Psi}\right)$$

- It looks like a regression problem but without inputs (we observe Y and want to infer X)
- We need to find suitable data *representation*
- Can we learn the underlying (generative) mapping?

Linear latent variable model

$$p(\mathbf{Y} \mid \mathbf{W}, \mathbf{X}) = \prod_{n=1}^{N} p(\mathbf{y}_n \mid \mathbf{W}, \mathbf{x}_n)$$

- It loo

(we ob

- We r
- Can w

The power of priors, $p(\mathbf{x})$

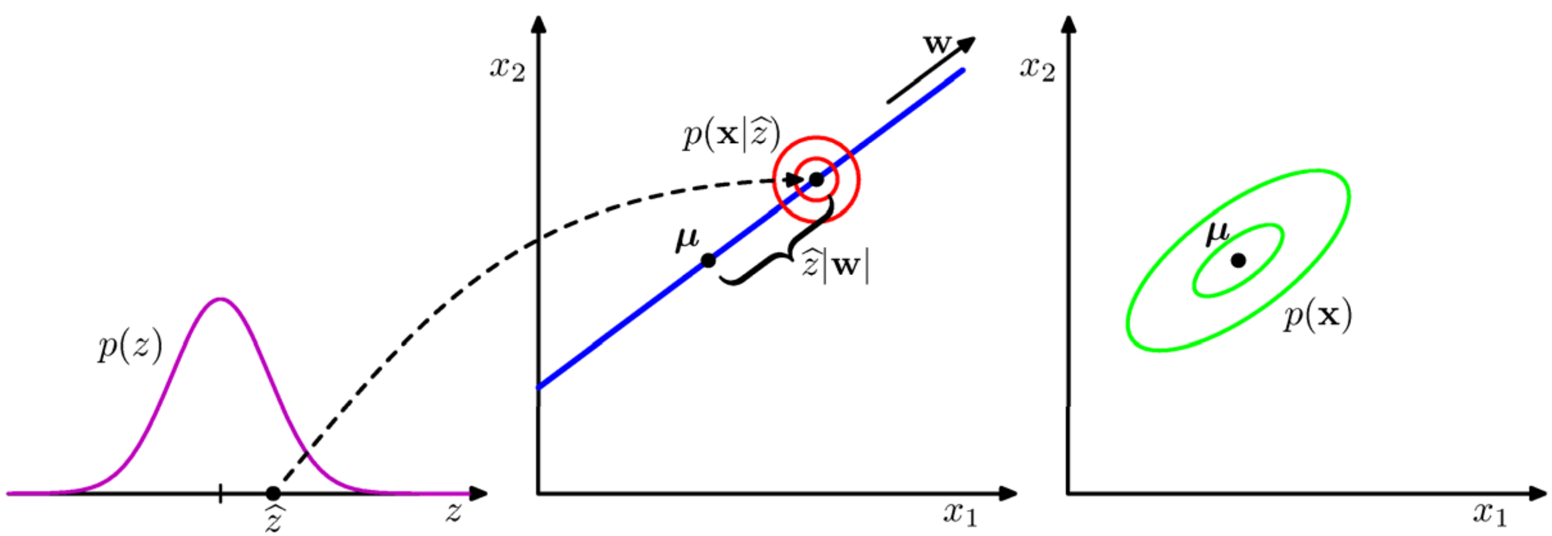
- encoding prior belief
- preference (choice out of many options)
- regularisation of the solution space

Factor analysis – general formulation

Generative mapping from the latent to the observed variable

$$\mathbf{y} = \mathbf{W}\mathbf{x} + \boldsymbol{\mu} + \varepsilon$$

$$\mathbf{y} \in \mathbb{R}^{D},\ \ \mathbf{x} \in \mathbb{R}^{M},\ \ \mathbf{W} \in \mathbb{R}^{D \times M}$$



$$p(x) = \mathcal{N}\left(x \,|\, \mu_0, \sigma_0\right)$$

Bishop 2006 (variable notation: $\mathbf{z}$ is $\mathbf{x}$, $\mathbf{x}$ is $\mathbf{y}$)

Factor analysis – general formulation

Generative mapping from the latent to the observed variable

$$\mathbf{y} = \mathbf{W}\mathbf{x} + \boldsymbol{\mu} + \varepsilon$$

$$\mathbf{y} \in \mathbb{R}^D, \ \mathbf{x} \in \mathbb{R}^M, \ \mathbf{W} \in \mathbb{R}^{D \times M}$$

factor loading matrix

Gaussian prior distribution over latent variables

$$p(\mathbf{x}) = \mathcal{N}\left(\mathbf{x} | \boldsymbol{\mu}_0, \Sigma_0\right)$$

Conditional distribution of the observed y conditioned on latent x

$$p(\mathbf{y} \mid \mathbf{W}, \mathbf{x}) = \mathcal{N}\left(\mathbf{W}\mathbf{x} + \boldsymbol{\mu}, \Psi\right)$$

Factor analysis – general formulation

Generative mapping from the latent to the observed variable

$$\mathbf{y} = \mathbf{W}\mathbf{x} + \boldsymbol{\mu} + \varepsilon$$

$$\mathbf{y} \in \mathbb{R}^{D},\ \ \mathbf{x} \in \mathbb{R}^{M},\ \ \mathbf{W} \in \mathbb{R}^{D\times M}$$

Gaussian

- The generative mapping is linear
- Assumption: $\boldsymbol{\Psi}$ is diagonal (to explain the correlation)
- Looks like regression but we have no inputs so there is need to specify **prior**

Conditiona x

$$p(\mathbf{y} \mid \mathbf{W}, \mathbf{x}) = \mathcal{N}\left(\mathbf{W}\mathbf{x} + \boldsymbol{\mu}, \boldsymbol{\Psi}\right)$$

Factor analysis – parameter learning

Marginal distribution

$$p(\mathbf{y}_n \mid \boldsymbol{\theta}) = \int \underbrace{p(\mathbf{y}_n \mid \mathbf{x}_n, \boldsymbol{\theta})}_{\text{conditional distr.}} \underbrace{p(\mathbf{x}_n)}_{\text{prior}} d\mathbf{x}_n = \int \mathcal{N}\left(\mathbf{W}\mathbf{x}_n + \boldsymbol{\mu}, \boldsymbol{\Psi}\right) \mathcal{N}\left(\boldsymbol{\mu}_0, \Sigma_0\right) d\mathbf{x}_n$$

Murphy (2002), ch.12

adapted from Carl Henrik Ek's lecture (DD2434, 2015)

Factor analysis – parameter learning

Marginal distribution

$$p(\mathbf{y}_n \mid \boldsymbol{\theta}) = \int \underbrace{p(\mathbf{y}_n \mid \mathbf{x}_n, \boldsymbol{\theta})}_{\text{conditional distr.}} \underbrace{p(\mathbf{x}_n)}_{\text{prior}} d\mathbf{x}_n = \int \mathcal{N}\left(\mathbf{W}\mathbf{x}_n + \boldsymbol{\mu}, \boldsymbol{\Psi}\right) \mathcal{N}\left(\boldsymbol{\mu}_0, \Sigma_0\right) d\mathbf{x}_n$$

- $\mathbf{X}$ and $\mathbf{W}$ are related
- here, we integrate out $\mathbf{X}$ (marginalisation)

Murphy (2002), ch.12

adapted from Carl Henrik Ek's lecture (DD2434, 2015)

Factor analysis – parameter learning

Marginal distribution

$$p(\mathbf{y}_n \mid \boldsymbol{\theta}) = \int \underbrace{p(\mathbf{y}_n \mid \mathbf{x}_n, \boldsymbol{\theta})}_{\text{conditional distr.}} \underbrace{p(\mathbf{x}_n)}_{\text{prior}} d\mathbf{x}_n = \int \mathcal{N}\left(\mathbf{W}\mathbf{x}_n + \boldsymbol{\mu}, \boldsymbol{\Psi}\right) \mathcal{N}\left(\boldsymbol{\mu}_0, \Sigma_0\right) d\mathbf{x}_n$$

$$\boxed{p(\mathbf{y}_i \mid \boldsymbol{\theta}) = \mathcal{N}\left(\mathbf{W}\boldsymbol{\mu}_0 + \boldsymbol{\mu}, \boldsymbol{\Psi} + \mathbf{W}\Sigma_0\mathbf{W}^{\mathrm{T}}\right)}$$

Murphy (2002), ch.12

Factor analysis – parameter learning

Marginal distribution

$$p(\mathbf{y}_n \mid \boldsymbol{\theta}) = \int \underbrace{p(\mathbf{y}_n \mid \mathbf{x}_n, \boldsymbol{\theta})}_{\text{conditional distr.}} \underbrace{p(\mathbf{x}_n)}_{\text{prior}} d\mathbf{x}_n = \int \mathcal{N}\left(\mathbf{W}\mathbf{x}_n + \boldsymbol{\mu}, \boldsymbol{\Psi}\right) \mathcal{N}\left(\boldsymbol{\mu}_0, \Sigma_0\right) d\mathbf{x}_n$$

$$p(\mathbf{y}_i \mid \boldsymbol{\theta}) = \mathcal{N}\left(\mathbf{W}\boldsymbol{\mu}_0 + \boldsymbol{\mu}, \boldsymbol{\Psi} + \mathbf{W}\Sigma_0\mathbf{W}^{\mathrm{T}}\right)$$

- if $\boldsymbol{\Sigma}_0 = \mathbf{I}$ and $\boldsymbol{\mu}_0 = \mathbf{0}$, i.e. $p(\mathbf{x}) = \mathcal{N}(0, \mathbf{I})$, then $p(\mathbf{y}_i) = \mathcal{N}\left(\boldsymbol{\mu}, \boldsymbol{\Psi} + \mathbf{W}\mathbf{W}^{\mathrm{T}}\right)$

Factor analysis – parameter learning

Marginal distribution

$$p(\mathbf{y}_n \mid \boldsymbol{\theta}) = \int \underbrace{p(\mathbf{y}_n \mid \mathbf{x}_n, \boldsymbol{\theta})}_{\text{conditional distr.}} \underbrace{p(\mathbf{x}_n)}_{\text{prior}} d\mathbf{x}_n = \int \mathcal{N}\left(\mathbf{W}\mathbf{x}_n + \boldsymbol{\mu}, \boldsymbol{\Psi}\right) \mathcal{N}\left(\boldsymbol{\mu}_0, \Sigma_0\right) d\mathbf{x}_n$$

$$p(\mathbf{y}_i \mid \boldsymbol{\theta}) = \mathcal{N}\left(\mathbf{W}\boldsymbol{\mu}_0 + \boldsymbol{\mu}, \boldsymbol{\Psi} + \mathbf{W}\Sigma_0\mathbf{W}^{\mathrm{T}}\right)$$

- if $\boldsymbol{\Sigma}_0 = \mathbf{I}$ and $\boldsymbol{\mu}_0 = \mathbf{0}$, then $p(\mathbf{y}_i) = \mathcal{N}\left(\boldsymbol{\mu}, \boldsymbol{\Psi} + \mathbf{W}\mathbf{W}^{\mathrm{T}}\right)$ ← rank of dim $\mathbf{X}$

A low rank parameterisation (density model) of the original $\mathbf{Y}$

Factor analysis – unidentifiability

If we try to rotate our weight matrix $\mathbf{W}$ by an orthogonal $\mathbf{R}$:

$$\tilde{\mathbf{W}} = \mathbf{W}\mathbf{R}$$

$$p(\mathbf{y}_n \mid \boldsymbol{\theta}) = \mathcal{N}\left(\boldsymbol{\mu}, \boldsymbol{\Psi} + \mathbf{W}\mathbf{R}\mathbf{R}^{\mathrm{T}}\mathbf{W}^{\mathrm{T}}\right)$$

$$= \mathcal{N}\left(\boldsymbol{\mu}, \boldsymbol{\Psi} + \mathbf{W}\mathbf{W}^{\mathrm{T}}\right)$$

Factor analysis – unidentifiability

If we try to rotate our weight matrix $\mathbf{W}$ by an orthogonal $\mathbf{R}$:

$$\tilde{\mathbf{W}} = \mathbf{W}\mathbf{R}$$

$$\begin{aligned} p(\mathbf{y}_n \mid \boldsymbol{\theta}) &= \mathcal{N}\left(\boldsymbol{\mu}, \boldsymbol{\Psi} + \mathbf{W}\mathbf{R}\mathbf{R}^{\mathrm{T}}\mathbf{W}^{\mathrm{T}}\right) \\ &= \mathcal{N}\left(\boldsymbol{\mu}, \boldsymbol{\Psi} + \mathbf{W}\mathbf{W}^{\mathrm{T}}\right) \end{aligned}$$

The marginal likelihood is *invariant* to a rotation

- ➢ no unique solution for $\mathbf{W}$ → we cannot uniquely identify latent factors
- ➢ interpretation becomes somewhat tricky
- ➢ it is possible to impose extra constraints on $\mathbf{W}$

Probabilistic PCA (PPCA)

Probabilistic formulation of PCA stems from factor analysis, if

1) Ψ is isotropic, i.e. $\Psi = \sigma^2 \mathbf{I}$ (the same noise variances for each variable)

2) factor loadings are orthogonal, i.e. $\mathbf{W}^\mathrm{T}\mathbf{W} = \mathbf{I}$

$$p(\mathbf{y} \mid \boldsymbol{\theta}) = \mathcal{N}\left(\boldsymbol{\mu}, \sigma^2 \mathbf{I} + \mathbf{W}\mathbf{W}^\mathrm{T}\right)$$

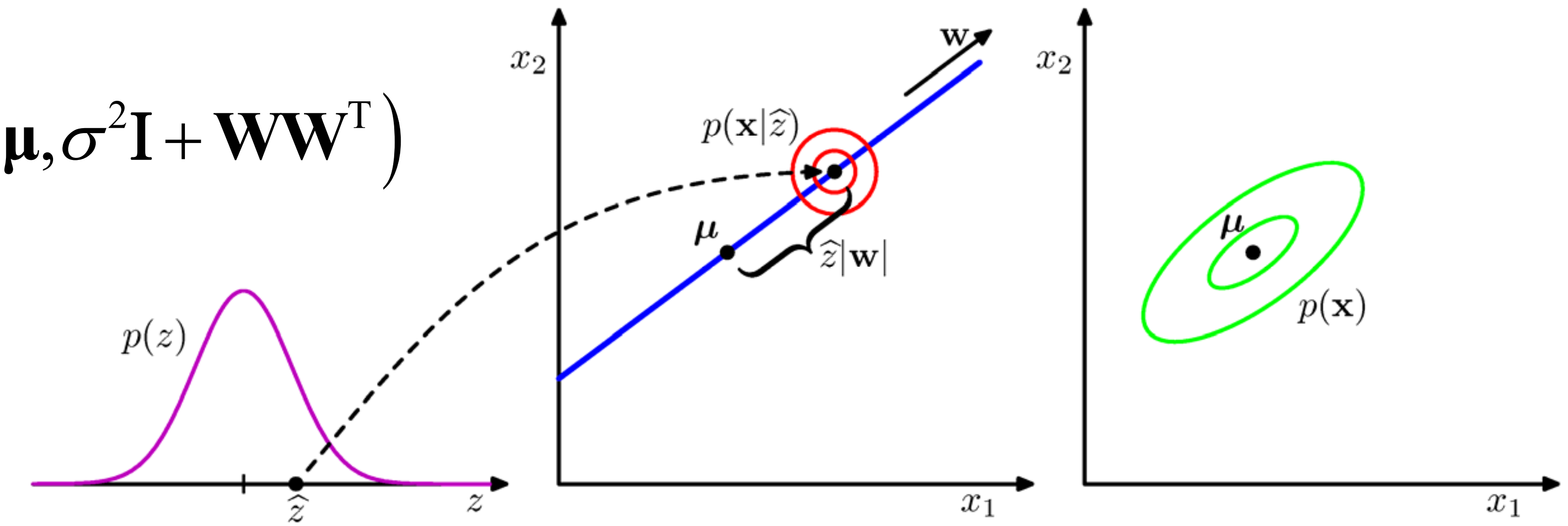
Probabilistic PCA (PPCA)

Probabilistic formulation of PCA stems from factor analysis, if

1) Ψ is isotropic, i.e. $\Psi = \sigma^2 \mathbf{I}$ (the same noise variances for each variable)

2) factor loadings are orthogonal, i.e. $\mathbf{W}^\mathrm{T}\mathbf{W} = \mathbf{I}$

$$p(\mathbf{y} \mid \boldsymbol{\theta}) = \mathcal{N}\left(\boldsymbol{\mu}, \sigma^2 \mathbf{I} + \mathbf{W}\mathbf{W}^\mathrm{T}\right)$$

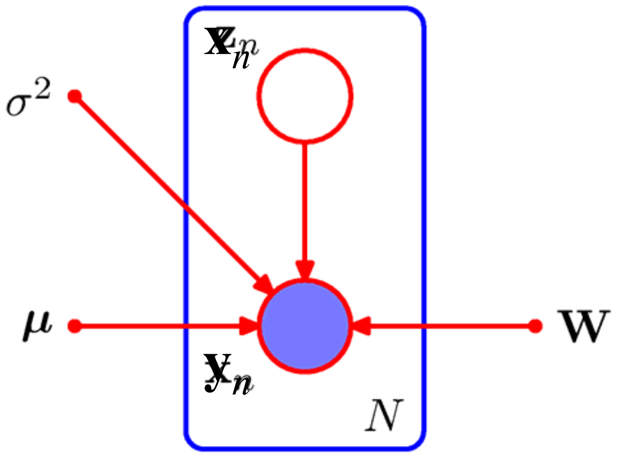


Bishop 2006 (variable notation: $\mathbf{z}$ is $\mathbf{x}$, $\mathbf{x}$ is $\mathbf{y}$)

Maximum Likelihood PCA

We first obtained marginal distribution by integrating out latent variable $\boldsymbol{x}$

$$p(\mathbf{y}_n \mid \boldsymbol{\theta}) = \int p(\mathbf{y}_n \mid \mathbf{x}_n, \boldsymbol{\theta}) p(\mathbf{x}_n) d\mathbf{x}_n = \mathcal{N}\left(\boldsymbol{\mu}, \sigma^2 \mathbf{I} + \mathbf{W}\mathbf{W}^{\mathrm{T}}\right)$$



Bishop 2006

Maximum Likelihood PCA

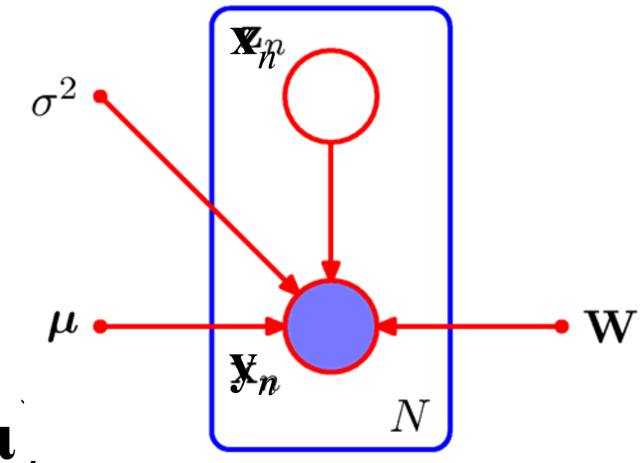
We first obtained marginal distribution by integrating out latent variable ***x***

$$p(\mathbf{y}_n \mid \boldsymbol{\theta}) = \int p(\mathbf{y}_n \mid \mathbf{x}_n, \boldsymbol{\theta}) p(\mathbf{x}_n) d\mathbf{x}_n = \mathcal{N}\left(\boldsymbol{\mu}, \underbrace{\sigma^2 \mathbf{I} + \mathbf{W}\mathbf{W}^{\mathrm{T}}}_{\mathbf{C}}\right)$$

Log-likelihood:

$$\log p(\mathbf{Y} \mid \boldsymbol{\mu}, \mathbf{W}, \sigma^2) = \sum_{n=1}^{N} \log p(\mathbf{y}_n \mid \boldsymbol{\mu}, \mathbf{W}, \sigma^2) =$$

$$= -\frac{ND}{2}\log(2\pi) - \frac{N}{2}\log|\mathbf{C}| - \frac{1}{2}\sum_{n=1}^{N}(\mathbf{y}_n - \boldsymbol{\mu})^{\mathrm{T}}\mathbf{C}^{-1}(\mathbf{y}_n - \boldsymbol{\mu})$$



Bishop 2006

Maximum Likelihood PCA

We first obtained marginal distribution by integrating out latent variable ***x***

$$p(\mathbf{y}_n \mid \boldsymbol{\theta}) = \int p(\mathbf{y}_n \mid \mathbf{x}_n, \boldsymbol{\theta}) p(\mathbf{x}_n) d\mathbf{x}_n = \mathcal{N}\left(\boldsymbol{\mu}, \underbrace{\sigma^2 \mathbf{I} + \mathbf{W}\mathbf{W}^{\mathrm{T}}}_{\mathbf{C}}\right)$$

Now we can select parameters based on the maximum likelihood principle: $\mathbf{W}_{\mathrm{ML}} = \arg\max_{\mathbf{W}} \left\{ \log p(\mathbf{Y} \mid \boldsymbol{\mu}, \mathbf{W}, \sigma^2) \right\}$

Bishop 2006

Maximum Likelihood PCA

We first obtained marginal distribution by integrating out latent variable $\boldsymbol{x}$

$$p(\mathbf{y}_n \mid \boldsymbol{\theta}) = \int p(\mathbf{y}_n \mid \mathbf{x}_n, \boldsymbol{\theta}) p(\mathbf{x}_n) d\mathbf{x}_n = \mathcal{N}\left(\boldsymbol{\mu}, \sigma^2 \mathbf{I} + \mathbf{W}\mathbf{W}^{\mathrm{T}}\right)$$

Now we can select parameters based on the maximum likelihood principle: $\mathbf{W}_{\mathrm{ML}} = \arg\max_{\mathbf{W}} \left\{ \log p(\mathbf{Y} \mid \boldsymbol{\mu}, \mathbf{W}, \sigma^2) \right\}$

$$\boxed{\begin{aligned} \mathbf{W}_{\mathrm{ML}} &= \mathbf{U}_M (\boldsymbol{\Lambda}_M - \sigma^2 \mathbf{I})^{1/2} \\ \boldsymbol{\mu} &= \overline{\mathbf{y}} \\ \sigma^2_{\mathrm{ML}} &= \frac{1}{D-M} \sum_{i=M+1}^{D} \lambda_i \end{aligned}}$$

$$\mathbf{U}_M : D \times M, \quad \boldsymbol{\Lambda}_M : M \times M$$

Tipping and Bishop, 1999

Maximum Likelihood PCA

Interpretations

$$\mathbf{W}_{\mathrm{ML}} = \mathbf{U}_M(\boldsymbol{\Lambda}_M - \sigma^2\mathbf{I})^{1/2}, \quad \boldsymbol{\mu} = \bar{\mathbf{y}}, \quad \sigma^2_{\mathrm{ML}} = \frac{1}{D-M}\sum_{i=M+1}^{D}\lambda_i$$

- Columns of $\mathbf{W}$ are PC eigenvec scaled by the variances $\lambda_i - \sigma^2$
- what is the variance in the direction $\mathbf{v}$ (unit vector), $\mathbf{v}^{\mathrm{T}}\mathbf{C}\mathbf{v}$?

Bishop, 2006

Maximum Likelihood PCA

Interpretations

$$\mathbf{W}_{\mathrm{ML}} = \mathbf{U}_M(\mathbf{\Lambda}_M - \sigma^2\mathbf{I})^{1/2}, \quad \boldsymbol{\mu} = \bar{\mathbf{y}}, \quad \sigma^2_{\mathrm{ML}} = \frac{1}{D-M}\sum_{i=M+1}^{D}\lambda_i$$

- Columns of $\mathbf{W}$ are PC eigenvec scaled by the variances $\lambda_i - \sigma^2$
- what is the variance in the direction $\mathbf{v}$ (unit vector), $\mathbf{v}^{\mathrm{T}}\mathbf{C}\mathbf{v}$?
  - if $\mathbf{v}$ is orthogonal to the principal space, then $\mathbf{v}^{\mathrm{T}}\mathbf{U}=0$ and $\mathbf{v}^{\mathrm{T}}\mathbf{C}\mathbf{v}=\sigma^2$
  - if $\mathbf{v} = \mathbf{u}_i$ then $\mathbf{v}^{\mathrm{T}}\mathbf{C}\mathbf{v}$ (variance) is $\lambda_i - \sigma^2 + \sigma^2 = \lambda_i$.

Bishop, 2006

Posterior over latent variables in PPCA

Posterior distribution ("reverse" mapping) can be obtained by applying Bayes' theorem and making calculus on Gaussian distr.

It is easier to make inverse if $\mathbf{\Psi} = \sigma^2\ \mathbf{I}$ (PPCA):

$$p(\mathbf{x}\,|\,\mathbf{y}) = \mathcal{N}\left(\underbrace{\left(\mathbf{W}^{\mathrm{T}}\mathbf{W}+\sigma^2\mathbf{I}\right)^{-1}\mathbf{W}^{\mathrm{T}}\left(\mathbf{y}-\boldsymbol{\mu}\right)}_{\text{posterior mean}},\underbrace{\sigma^2\left(\mathbf{W}^{\mathrm{T}}\mathbf{W}+\sigma^2\mathbf{I}\right)^{-1}}_{\text{posterior covariance}}\right)$$

Posterior over latent variables in PPCA

Posterior distribution (“reverse” mapping) can be obtained by applying Bayes’ theorem and making calculus on Gaussian distr.

It is easier to make inverse if $\mathbf{\Psi} = \sigma^2 \mathbf{I}$ (PPCA):

$$p(\mathbf{x} \mid \mathbf{y}) = \mathcal{N}\left(\left(\mathbf{W}^{\mathrm{T}}\mathbf{W} + \sigma^2\mathbf{I}\right)^{-1} \mathbf{W}^{\mathrm{T}}(\mathbf{y} - \boldsymbol{\mu}), \sigma^2\left(\mathbf{W}^{\mathrm{T}}\mathbf{W} + \sigma^2\mathbf{I}\right)^{-1}\right)$$

So, the point targeted (projected to) in data from latent space is

$$\mathbf{W}\mathbb{E}[\mathbf{x} \mid \mathbf{y}] + \boldsymbol{\mu}$$

Posterior over latent variables in PPCA

Posterior distribution ("reverse" mapping) can be obtained by applying Bayes' theorem and making calculus on Gaussian distr.

It is easier to make inverse if $\mathbf{\Psi} = \sigma^2\,\mathbf{I}$ (PPCA):

$$p(\mathbf{x}\,|\,\mathbf{y}) = \mathcal{N}\left(\left(\mathbf{W}^{\mathrm{T}}\mathbf{W}+\sigma^2\mathbf{I}\right)^{-1}\mathbf{W}^{\mathrm{T}}\left(\mathbf{y}-\boldsymbol{\mu}\right), \sigma^2\left(\mathbf{W}^{\mathrm{T}}\mathbf{W}+\sigma^2\mathbf{I}\right)^{-1}\right)$$

For $\mathbf{W}=\mathbf{W}_{\mathrm{ML}}$ and in the limit $\sigma^2 \to 0$, the posterior mean becomes

$$\left(\mathbf{W}_{\mathrm{ML}}^{\mathrm{T}}\mathbf{W}_{\mathrm{ML}}\right)^{-1}\mathbf{W}_{\mathrm{ML}}^{\mathrm{T}}\left(\mathbf{y}-\overline{\mathbf{y}}\right)$$

Posterior over latent variables in PPCA

Posterior distribution ("reverse" mapping) can be obtained by applying Bayes' theorem and making calculus on Gaussian distr.

It is easier to make inverse if $\mathbf{\Psi} = \sigma^2\, \mathbf{I}$ (PPCA):

$$p(\mathbf{x}\,|\,\mathbf{y}) = \mathcal{N}\left(\left(\mathbf{W}^{\mathrm{T}}\mathbf{W}+\sigma^2\mathbf{I}\right)^{-1}\mathbf{W}^{\mathrm{T}}(\mathbf{y}-\boldsymbol{\mu}), \sigma^2\left(\mathbf{W}^{\mathrm{T}}\mathbf{W}+\sigma^2\mathbf{I}\right)^{-1}\right)$$

For $\mathbf{W}=\mathbf{W}_{\mathrm{ML}}$ and in the limit $\sigma^2 \to 0$, the posterior mean becomes

$$\left(\mathbf{W}_{\mathrm{ML}}^{\mathrm{T}}\mathbf{W}_{\mathrm{ML}}\right)^{-1}\mathbf{W}_{\mathrm{ML}}^{\mathrm{T}}(\mathbf{y}-\bar{\mathbf{y}}) \to \mathbf{W}_{\mathrm{ML}}^{\mathrm{T}}(\mathbf{y}-\bar{\mathbf{y}})$$

if $\mathbf{W}$ is orthogonal

……. and the posterior covariance = 0.

Posterior over latent variables in PPCA

Posterior distribution ("reverse" mapping) can be obtained by applying Bayes' theorem and making calculus on Gaussian distr.

It is easier to make inverse if $\mathbf{\Psi} = \sigma^2\,\mathbf{I}$ (PPCA):

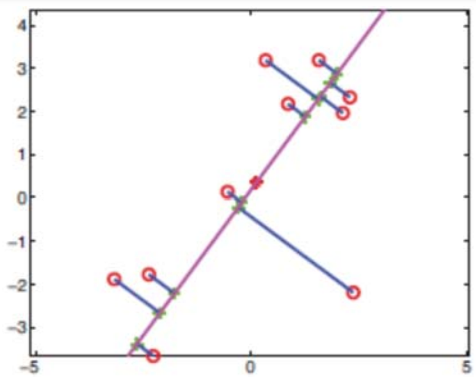
p

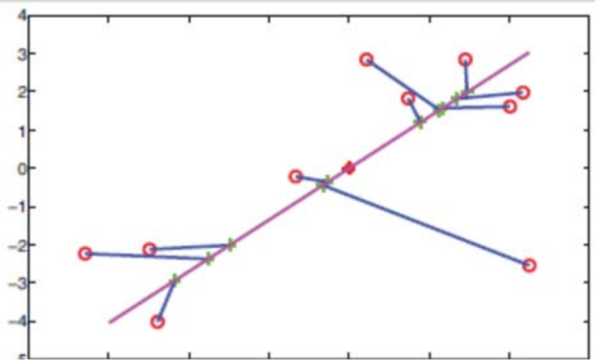
Fo

an

- For $\sigma^2 \to 0$ we recover the *classical PCA* formulation!
- For $\sigma^2 > 0$ we obtain PPCA (*sensible PCA*)
  - the posterior mean is no longer an orthogonal projection
  - it is shrunk towards the prior mean
  - the reconstruction is closer to the overall data mean $\bar{\mathbf{y}} = \boldsymbol{\mu}$

Posterior over latent variables in PPCA





Fo

an

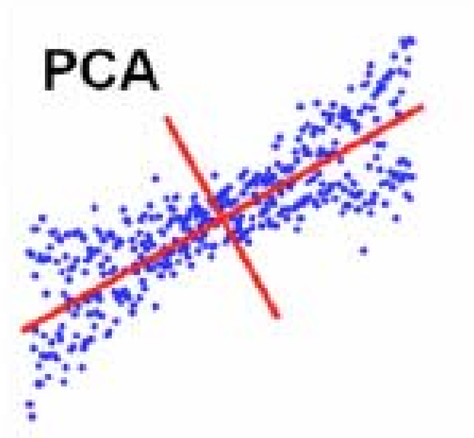
- For $\sigma^2 \to 0$ we recover the *classical PCA* formulation!
- For $\sigma^2 > 0$ we obtain PPCA (*sensible PCA*)
  - the posterior mean is no longer an orthogonal projection
  - it is shrunk towards the prior mean
  - the reconstruction is closer to the overall data mean $\overline{\mathbf{y}} = \boldsymbol{\mu}$

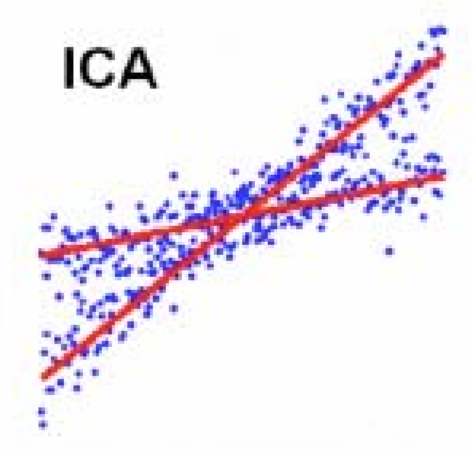
Independent component analysis (ICA)

- Cocktail party problem (blind source separation)
- Factorisation of the latent variables distribution

$$p(\mathbf{x}) = \prod_{j=1}^{M} p(x_j)$$

- Non-gaussian distributions of different sources
- The prior determines different models





Summary

- Graphical models – an intuitive way of representing and visualising the relationships between many variables
- Implicit representation and degrees of freedom
- Generative models
- Priors as preference
- Factor analysis is a linear continuous latent variable model
- PPCA is a factor analysis with special conditions: isotropic covariance matrix and orthogonality of W
- PPCA reduces to classical PCA formulation with $\sigma^2 \to 0$

Recommended additional reading

On PPCA

M.E. Tipping and C.M. Bishop. (1999) Probabilistic principal component analysis. *Journal of the Royal Statistical Society*, Series B 21(3), 611–622.6.

K.P. Murphy (2012) *Machine Learning : Probabilistic Perspective*. Chapter 12. MIT Press.

On GP-LVMs

Neil D Lawrence. (2005) Probabilistic non-linear principal component analysis with Gaussian process latent variable models". *The Journal of Machine Learning Research* 6